Avaliação com itens abertos: validade, confiabilidade, comparabilidade e justiça

Sônia Ferreira Lopes Toffoli[I]
Dalton Francisco de Andrade[II]
Antonio Cezar Bornia[II]
Gladys Quevedo-Camargo[III]

Resumo

As avaliações em larga escala, dependendo da área na qual estão sendo aplicadas, são responsáveis por orientar decisões importantes. Nos exames educacionais, os objetivos podem estar direcionados para as diferenças individuais, monitorando o desempenho dos estudantes em diversas situações, como também para a apreciação de programas ou de projetos educacionais, subsidiando ou justificando alguma ação na esfera política. A validade das medidas e suas interpretações são de suma importância, com consequências que podem afetar a população envolvida e até mesmo a sociedade. As questões consideradas fundamentais para uma avaliação em larga escala eficiente consistem em validade, confiabilidade, comparabilidade e justiça. Esses termos devem ser considerados sempre que decisões de valores são tomadas com base nas avaliações. São feitas considerações sobre os conceitos de validade e de confiabilidade e a relação existente entre eles. A comparação entre avaliações com itens abertos consiste atualmente em uma das questões de maior preocupação para os especialistas, fato esse provocado pela intensificação da utilização de matrizes comuns de referência desenvolvidas para orientar os currículos em todos os níveis de ensino em diversas nações. Discute-se também a justiça nas avaliações, que está relacionada com a igualdade de condições a todos os seus participantes. Uma avaliação de qualidade deve permitir às pessoas oportunidades de respostas que assegurem inferências corretas sobre seu desempenho em relação ao construto medido. O objetivo deste trabalho é descrever as principais teorias presentes nas avaliações em larga escala, fornecendo subsídios para uma correta interpretação dos conceitos envolvidos em seus processos.

Palavras-chave

Avaliação em larga escala – Validade – Confiabilidade – Comparabilidade – Justiça.

I- Universidade Estadual de Londrina, PR, Brasil.
Contato: sonialopes@uel.br
II- Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil.
Contatos: dandrade@inf.ufsc.br; cezar@inf.ufsc.br
III- Universidade de Brasília DF, Brasil.
Contato: gladys@unb.br

Assessment with construct-response items: *validity, reliability, comparability, and fairness*

Sônia Ferreira Lopes Toffoli[I]
Dalton Francisco de Andrade[II]
Antonio Cezar Bornia[II]
Gladys Quevedo-Camargo[III]

Abstract

Large-scale assessments may guide important decisions depending on the area in which they are applied. In educational exams, objectives may focus on individual differences, monitoring the performance of students in different contexts, as well as on the assessment of educational programs or projects, supporting or justifying actions in the political sphere. The validity of the measures and their interpretation are of paramount importance, as their consequences may affect the population involved and even the whole society. The key issues for large-scale assessment are validity, reliability, comparability, and fairness. These terms should be considered whenever value decisions are made based on the assessments. This article discusses the concepts of validity and reliability, as well as the relationship between them. The comparison of assessments with construct-response items is currently an issue of great concern to experts, due to the increased use of shared reference matrices developed to guide curricula at all educational levels in several nations. This article also discusses fairness in evaluations, which is related to the requirement to ensure equal conditions to all participants. Quality assessment should provide all with opportunities for responses which ensure correct inferences about their performance in relation to the construct measured. The aim of this article is to describe the main theories present in large-scale assessments, providing information for the correct interpretation of the concepts involved in their processes.

Keywords

Large-scale assessment – Validity – Reliability – Comparability – Fairness.

I- Universidade Estadual de Londrina, PR, Brasil.
Contact: sonialopes@uel.br
II- Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil.
Contact: dandrade@inf.ufsc.br; cezar@inf.ufsc.br
III- Universidade de Brasília DF, Brasil.
Contact: gladys@unb.br

Introdução

Os processos avaliativos possuem diferentes objetivos, como classificação de candidatos com a finalidade de aprovação para um emprego ou vaga de escola, determinação do grau de habilidade para uma atividade específica, entre outras. As informações provenientes das avaliações auxiliam as decisões pessoais ou da esfera pública. É por essa razão que as avaliações devem ser confiáveis.

No âmbito educacional, as avaliações em larga escala são aquelas elaboradas e aplicadas para um grande número de pessoas e possuem objetivos diferentes das avaliações aplicadas dentro da sala de aula. Elas exercem uma forte influência sobre as políticas educacionais e os currículos nos diversos níveis de ensino em todo o mundo. Portanto, é evidente a importância de examinar as variáveis envolvidas na construção, aplicação e pontuação desses exames (BEHIZADEH; ENGELHARD, 2011; SCARAMUCCI, 2011; BESSA, 2007).

São frequentemente designados na literatura como avaliações de desempenho os testes que necessitam do julgamento de avaliadores especialistas (ENGELHARD, 2013; ECKES, 2011). Essa classe de avaliações é utilizada em uma variedade de áreas, por exemplo, em competições esportivas em que os atletas executam uma série de exercícios na presença de especialistas que julgam a qualidade da apresentação com base em critérios preestabelecidos. Os testes com itens de respostas abertas, para os quais o indivíduo pode elaborar uma resposta escrita, são também avaliações de desempenho, pois necessitam de especialistas para corrigir as respostas. Nessa categoria, enquadram-se também as redações dos vestibulares e de outros concursos, testes orais e entrevistas para seleção.

Para maior simplicidade e uniformidade na denominação, as avaliações com itens de respostas abertas serão referidas neste trabalho simplesmente como avaliações com itens abertos.

No Brasil, as avaliações com itens abertos mais utilizadas nos exames educacionais e de seleção em larga escala são as avaliações das disciplinas do ensino médio de exames vestibulares de algumas instituições e a prova de redação também presente nos exames vestibulares, no ENEM (Exame Nacional do Ensino Médio) e em alguns concursos públicos e privados para suprir vagas de trabalho. Além desses exemplos, outras avaliações com itens abertos não são muito comuns no Brasil. Para exemplificar, podem-se citar algumas poucas edições do Sistema de Avaliação da Educação Básica (SAEB) e do Programa Internacional de Avaliação de Alunos (PISA) que utilizam itens abertos para avaliar a linguagem e a matemática (KLEIN; FONTANIVE, 2009).

A avaliação com itens abertos tem sido considerada uma área problemática desde a década de 1920, quando os primeiros estudos avaliavam aspectos limitados do desenvolvimento e da fluência na escrita. Entretanto, somente a partir da década de 1950, professores e pesquisadores intensificaram as buscas por métodos capazes de produzir validade e confiabilidade para essas avaliações[1] (BEHIZADEH; ENGELHARD, 2011; HAMP-LYONS, 2004; HUOT, 1990).

A qualidade das avaliações, de certa forma, é comumente alcançada e relatada em termos desses conceitos e, nas últimas décadas, tem havido uma alternância quanto à sua importância. Dependendo dos pesquisadores e do período considerado, um destes conceitos foi privilegiado em detrimento do outro (YANCEY, 1999).

Além da validade e da confiabilidade, outras duas questões são consideradas fundamentais para uma avaliação eficiente: a comparabilidade entre avaliações distintas e a justiça. A primeira diz respeito à possibilidade de comparar resultados provenientes de diferentes administrações de uma avaliação, o que é essencial para a obtenção de indicadores de tendências de desempenho acadêmico, importantes tanto para a utilização nas

1- Os conceitos de validade e de confiabilidade serão discutidos no decorrer deste trabalho.

esferas políticas quanto em polos educacionais (ELLIOTT, 2013; HAERTEL; LINN, 1996).

A comparação entre testes que utilizam itens de múltipla escolha encontra-se estabelecida e vem sendo amplamente utilizada em avaliações em muitos países. Como exemplos brasileiros estão o ENEM e o SAEB; por outro lado, a comparação entre as avaliações com itens abertos é mais difícil de ser alcançada. Entre os motivos está o fato de os testes serem menos padronizados e envolverem menos itens e de serem, na maioria, pontuados por avaliadores humanos (ECKES, 2011; ENGELHARD, 2013). Uma avaliação em larga escala internacional que proporciona comparações é o PISA. A finalidade é confrontar as proficiências dos alunos dos países participantes para produzir indicadores que contribuam para a discussão da qualidade da educação, de modo a subsidiar políticas de melhoria do ensino básico.

O termo justiça no desenvolvimento e utilização das avaliações educacionais refere-se à questão da tendenciosidade. Esse conceito está relacionado com a equidade do teste ou a possibilidade de garantir oportunidades iguais a todos os participantes (ETS, 2009). A justiça (equidade) baseada apenas em comparações diversas entre grupos, avaliações etc. é frequentemente rejeitada pela comunidade científica, uma vez que as diferenciações podem ocorrer devido a fatores que invalidam as análises e os julgamentos dos resultados observados. Os problemas relacionados com a justiça nas avaliações envolvem questões de política educacional que refletem tensões sociais e interesses de natureza diversa (BESSA, 2007).

A finalidade deste artigo é destacar os procedimentos essenciais necessários para que as avaliações em larga escala com itens abertos sejam desenvolvidas de acordo com padrões atuais de qualidade, que envolvem questões essenciais para a avaliação, a saber, validade, confiabilidade, comparabilidade e justiça.

Segundo Messick (1996), os conceitos de validade, confiabilidade, comparabilidade e justiça, fundamentais para uma avaliação eficiente, não se resumem apenas a princípios de medição; são valores sociais com significado, e devem ser considerados sempre que decisões de valores são tomadas com base nas avaliações. Esses conceitos são descritos a seguir.

Validade

O conceito de validade vem sendo proposto e modificado desde os anos 1920, consistindo, juntamente com o de confiabilidade, provavelmente os conceitos mais polêmicos da área de avaliação. Iniciou-se com a definição proposta por Kelley (1927, p. 14): "um teste é válido se mede o que pretende medir". Apesar dessa definição ter sido publicada pela primeira vez há quase um século, ela ainda é muito utilizada (HAMP-LYONS, 2011; BEHIZADEH; ENGELHARD; 2011; PASQUALI, 2007; YANCEY, 1999). Esse conceito de validade é centro de muitas críticas. A principal delas é que, desse modo, a validade consiste em uma característica ou qualidade do teste, não levando em consideração o significado dos escores ou as consequências sociais e políticas do uso dos resultados (SCARAMUCCI, 2011). Tal crítica data de 1955, quando Cronbach e Meehl escreveram que a validade não se refere apenas a uma propriedade do teste, mas também às interpretações da sua pontuação (BORSBOOM; MELLENBERG; VAN HEERDEN, 2004).

Mais de trinta anos depois, Messick, em 1989, propôs o conceito de validade, considerado, nos dias de hoje, o modo moderno de entender a validade:

> A validade é um julgamento avaliativo integrado do grau em que as evidências empíricas e teóricas sustentam a adequação e a qualidade das inferências e ações com base nos resultados de testes ou em outros meios de avaliação (MESSICK, 1989, p. 13, tradução nossa)[2].

2- No original em inglês, "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support

Assim, esse conceito de validade consiste em saber se as interpretações e ações sobre os resultados dos testes são justificadas, tanto com base nas evidências científicas quanto nas consequências sociais e éticas da utilização de teste. Assim, a teoria da validade passou gradualmente a tratar todas as questões relacionadas aos testes e a integrar todas em uma única definição (PASQUALI, 2007; BORSBOOM; MELLENBERG; VAN HEERDEN, 2004; BESSA, 2007).

Essa preocupação com as interpretações e a utilização dos resultados dos testes levanta a questão da legitimidade da incorporação das consequências do teste para sua validação. Uma grande controvérsia consiste na questão da responsabilidade dos desenvolvedores do teste pela sua utilização. As discussões sobre esse assunto resultaram na criação do termo validade consequencial, e também num aumento das discussões sobre o impacto ou efeito retroativo[3] das avaliações em diferentes agentes envolvidos no processo avaliativo.

Esse conceito de validade proposto por Messick (1989) passou a ser utilizado por pesquisadores (SCARAMUCCI, 2011; MOSKAL; LEYDENS, 2000; CHAPELLE, 1999). Mesmo assim, ainda não foi estabelecido o consenso. Pasqualli (2007) e Borsboom, Mellenberg e van Heerden (2004), entre outros, consideram que a teoria de validade assim estabelecida torna o processo confuso para os responsáveis pela elaboração de testes, pois resulta em uma sensação de que tudo que se refere aos testes é relevante, tornando difíceis as decisões quanto aos aspectos de maior ou de menor importância.

É por esse motivo e também pela simplicidade que, ainda hoje, o conceito formulado por Kelley (1927), "um teste é válido se mede o que foi proposto a medir", é utilizado e considerado correto por muitos pesquisadores de renome, como Pasquali (2007), Borsboom, Mellenberg e van Heerden (2004) e Li (2003).

Vários tipos ou visões de validade existem na literatura e não há consenso quanto aos nomes, definições e métodos utilizados para a medida da característica pretendida (SCARAMUCCI, 2011; PASQUALLI, 2007; JONSSON; SVINGBY, 2007). Alguns exemplos são: validade de construto, validade de conteúdo, validade de face, validade de critério, entre outros.

A validade de construto refere-se ao grau em que o instrumento de pontuação é capaz de distinguir as habilidades (construtos) que se propõe a averiguar (PASQUALI, 2007). A validade de conteúdo refere-se à amostra do conteúdo abordado no teste, e se esse é relevante e representativo de todo o universo de conteúdo (PASQUALI, 2007; SCARAMUCCI, 2011). A validade de face ou aparente consiste em se ter os conteúdos de um teste analisados por especialistas para determinar se eles são apropriados (PASQUALI, 2007). A validade de critério diz respeito a um critério externo ao teste, como a previsão de sucesso do participante no futuro ao desempenhar tarefas correlacionadas ao teste (KANE, 2012; SCARAMUCCI, 2011; PASQUALLI, 2007).

Para Moskal e Leydens (2000), um instrumento é validado pelo processo de acumulação de evidências que suportam a adequação das inferências feitas das respostas dos alunos ao serem avaliados, e normalmente três tipos de evidências são requeridas: a validade de conteúdo, de construto e de critério.

Chapelle (1999) e Jonsson e Svingby (2007) consideram que a validade refere-se à construção de um conceito unificador que incorpora os diferentes aspectos de validade. A validade de construto ocupa um lugar central, isto é, a validade de conteúdo e a validade de critério podem ser utilizadas como evidências para estabelecer uma imagem mais completa da validade.

No Brasil, pouco se sabe sobre a aceitação ou uso de um ou outro conceito de validade, principalmente porque as informações provenientes das instituições promotoras das principais avaliações em larga escala são raras.

the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (MESSICK, 1989, p. 13).

3- Sobre efeito retroativo, ver Quevedo-Camargo (2014), Scaramucci (2004) e Alderson e Wall (1993).

O psicometrista brasileiro Pasqualli, em artigo publicado em 2007, é claramente contrário ao conceito de validade proposto por Messick (1989), e assegura que os instrumentos de medidas são desenvolvidos com a finalidade de avaliar traços latentes (construtos); desse modo, a qualidade do teste deve ser dada em relação à medida obtida do construto, objetivo da aplicação do mesmo. Já a pesquisadora Scaramucci, parece defender a utilização do novo conceito de validade. Entretanto, faz algumas considerações quando afirma que

> [...] somente podemos contabilizar como consequências de uma avaliação aquelas que realmente puderem ser atribuídas ao teste e a nada mais. O que nos coloca um problema de difícil solução: se as características desse instrumento interagem com outras forças na sociedade, como podemos separar o que é efeito de um teste de outras ações, tais como bom ensino, por exemplo? (SCARAMUCCI, 2011, p. 116).

Bessa (2007) ressalta que o novo conceito de validade proporcionou análises mais claras sobre o conteúdo dos testes e que as avaliações passaram a ser projetadas minuciosamente, de modo a contemplar a congruência dos elementos que as compõem, contribuindo assim, ao argumento de validade.

Embora existam poucas informações sobre esse debate no Brasil, pode-se intuir a não unanimidade para a adoção do novo conceito de validade, que inclui, como parte das evidências para a validação, as consequências das avaliações no ensino e aprendizagem, na política educacional, entre outras, apesar do reconhecimento por parte dos pesquisadores da importância dessas evidências.

Confiabilidade

O conceito de confiabilidade refere-se à consistência dos escores de avaliação. Isso significa que é esperado que um indivíduo alcance o mesmo resultado independentemente da ocasião em que este respondeu ao teste.

A confiabilidade de um teste, na visão de psicometristas como Guilford (1954) e Lord, Novick e Bimbaum (1968), é definida como uma correlação entre o escore verdadeiro e o observado. Assim, a confiabilidade é dependente do conceito de erro de medição, uma vez que o escore observado é o escore verdadeiro acrescido de um erro (LI, 2003). O grau de erro de medição está inversamente relacionado com o grau de confiabilidade: quanto maior for o erro, menor é a confiabilidade do teste, e vice-versa. Inferências baseadas em resultados de testes com confiabilidade pequena ou zero terão pouco valor, porque as pontuações são resultados de medição com erro muito alto ou total. Por essa razão, a validade é dependente da confiabilidade (SLOMP; FUITE, 2005).

Os resultados das avaliações de desempenho, especialmente da escrita, não dependem apenas do nível de habilidade dos indivíduos quanto ao construto avaliado e da dificuldade das tarefas; dependem também da severidade dos avaliadores que julgam os desempenhos e da estrutura da escala de classificação.

Um dos principais problemas nessas avaliações é a pontuação de um mesmo desempenho com graus diferentes de severidade. Quando existem vários avaliadores, seria ideal se todos atribuíssem exatamente a mesma pontuação para os mesmos desempenhos observados, condição principal para se ter confiabilidade de pontuação. Entretanto, são muitos os fatores que podem causar variabilidade nessas pontuações, especialmente quando se trata dos testes com itens abertos. As características pessoais dos avaliadores, como cultura, experiências, expectativas, estilo de correção, entre outras, podem influenciar a pontuação das tarefas. Esses fatores podem ser tão importantes para a pontuação quanto a qualidade da resposta escrita pelo examinando (ENGELHARD, 2013; ECKES, 2011; MYFORD; WOLFE, 2004).

Outra questão que interfere na obtenção de bons índices de confiabilidade é a tendência

dos avaliadores em julgamentos sistemáticos dos desempenhos avaliados. Essas tendências são comportamentos frequentemente relatados nas pesquisas e são consideradas componentes geradores de erros importantes na pontuação de tarefas escritas (ENGELHARD, 2013; ECKES, 2011; MYFORD; WOLFE, 2004). Alguns dos efeitos são: efeito da severidade, que é a tendência a avaliar de maneira muito exigente ou muito branda em comparação com a pontuação atribuída por outros avaliadores ou em comparação com classificações preestabelecidas como referência; o efeito halo, que ocorre quando os avaliadores não conseguem distinguir entre um número de categorias conceitualmente distintas e avaliam o desempenho da pessoa com base em uma impressão geral, fazendo com que diferentes desempenhos possam obter a mesma pontuação; o efeito de tendência central, caracterizado por estabelecer classificações perto do ponto médio da escala, evitando classificações nos extremos da escala; o efeito de aleatoriedade, que é a tendência do avaliador a aplicar uma ou mais categorias da escala de maneira inconsistente com o modo como os outros avaliadores aplicam a mesma escala. O avaliador que possui essa tendência é demasiadamente inconsistente no uso da escala, apresentando maior variabilidade aleatória que o esperado na avaliação (MYFORD; WOLFE, 2004).

Duas formas de confiabilidade normalmente são consideradas em avaliações: a interavaliador, em que os avaliadores concordam entre si em suas notas, e a intra-avaliador, em que cada avaliador atribui a mesma pontuação para um desempenho avaliado em ocasiões distintas (MOSKAL; LEYDENS, 2000).

A confiabilidade interavaliador ou confiabilidade entre examinadores independentes (sem discussão ou colaboração) é considerada atualmente a característica mais importante da avaliação com itens abertos. No entanto, é uma condição necessária, mas não suficiente para a validade. Isso significa que, sem um nível suficiente de acordo entre avaliadores, um procedimento de avaliação escrita não pode ser válido (HUOT, 1996). Por esse motivo, historicamente, a confiabilidade tem dominado a literatura sobre as avaliações escritas, pois só após o desenvolvimento de procedimentos e critérios de pontuação e de treinamento de avaliadores é que a avaliação da escrita tornou-se psicometricamente viável (HUOT, 1996).

Para garantir melhores índices de confiabilidade, geralmente são utilizados critérios de pontuação na forma de rubricas. Elas são esquemas descritivos desenvolvidos com a finalidade de detalhar como a pontuação deve ser atribuída, orientando as análises dos produtos ou processos elaborados pelos participantes da avaliação (MOSKAL; LEYDENS, 2000). Assim, as rubricas são utilizadas para diminuir a subjetividade na atribuição de notas e guiar os avaliadores para que alcancem uma pontuação confiável no julgamento de uma habilidade, proporcionando resultados melhores, o que é desejado para os avaliadores e para os participantes, independentemente de a avaliação ocorrer em ambiente escolar ou em larga escala (ECKES, 2011; JONSSON; SVINGBY, 2007).

O ENEM divulga anualmente um guia para informar ao participante a metodologia utilizada para a correção da redação e discorre sobre as competências avaliadas, as rubricas de pontuação e a escala. O PISA disponibiliza os critérios para a correção de alguns itens liberados para a divulgação, isso porque os itens de suas provas fazem parte de um banco de itens e podem ser repetidos em outras edições da avaliação. Esses documentos podem ser conferidos e são exemplos importantes de critérios de pontuação aplicados em avaliações em larga escala no Brasil (BRASIL, 2013, 2014).

Comparabilidade

A comparação entre avaliações com itens abertos consiste atualmente em uma das questões de maior preocupação para os especialistas, devido à intensificação da utilização de matrizes comuns de referência desenvolvidas para orientar os currículos em

todos os níveis de ensino, em países da Europa, nos Estados Unidos, na Austrália, no Brasil, entre outros (HAMP-LYONS, 2004).

A comparabilidade diz respeito à validade das inferências sobre comparações que são feitas com base em resultados de avaliações. Os estudos sobre comparabilidade entre avaliações são comuns na maioria dos países como, por exemplo, Fontanive et al (2010) no Brasil, Jeferry (2009) nos Estados Unidos e Elliott (2011) na Inglaterra. Entretanto, no Reino Unido, há um esforço por parte dos pesquisadores em busca de uma definição de comparabilidade que seja aceita e utilizada em todas as situações, uma vez que a comparabilidade é uma área cercada por suposições e por disputas metodológicas e considerada por alguns um terreno estéril, fadado ao fracasso (ELLIOT, 2013).

No campo das avaliações, quando se fala em comparabilidade, o pensamento incide sobre os processos pelos quais os resultados dos testes são traduzidos em normas ou padrões interpretáveis. Diferentes definições de comparabilidade podem ser encontradas na literatura, assim como são várias as técnicas para a sua determinação ou acompanhamento. Basicamente, três abordagens são comumente utilizadas para julgamento do padrão especificado para os exames: (1) em termos de critérios de desempenho, a qual considera apenas as características do teste; (2) em termos de normas estatísticas, a qual leva em conta o desempenho dos examinandos de uma população; e (3) em termos do desempenho em relação ao construto comum (COE, 2010). A ideia de um estudo empírico definitivo sobre a comparabilidade é um desafio que dependerá fortemente da definição adotada para apoiar a validade da técnica particular utilizada para o seu monitoramento (COE, 2010; POLLITT; AHMED; CRISP, 2007).

Embora existam diferentes formas e concepções de comparabilidade, elas geralmente podem ser enquadradas em uma das três abordagens mencionadas. Por meio das definições e discussões sobre cada uma dessas abordagens, são estabelecidas também as consequências e implicações que essas concepções têm sobre os termos *dificuldade* e *padrão,* que muitas vezes, são usados com objetivos diferentes.

A concepção de *comparabilidade de desempenho* é baseada no julgamento dos padrões de um exame, considerando apenas os resultados observados. Nesse contexto, a *dificuldade* é estabelecida por análises das exigências feitas ao participante da avaliação. As notas provenientes de dois exames diferentes são tomadas como equivalentes quando representam níveis comparáveis de desafio, complexidade ou habilidades necessárias. Nessa concepção, o julgamento do desempenho padrão é feito de acordo com a natureza e o nível de exigências que são necessárias para alcançá-lo, independentemente do número de indivíduos que o alcançou (COE, 2010).

A principal limitação dessa abordagem é que, muitas vezes, não é possível observar diretamente as habilidades demandadas e os níveis de desafio que devem ser deduzidos a partir dos fenômenos apresentados. Por isso, frequentemente são necessários conhecimentos ou suposições sobre o contexto da avaliação, o que torna as inferências problemáticas e complexas (POLLITT; AHMED; CRISP, 2007).

Segundo a concepção de *comparabilidade estatística*, o padrão depende essencialmente da probabilidade de ele ser atingido. Embora possam ser considerados outros fatores, como as exigências do exame ou a qualidade do desempenho, a comparabilidade estatística deve ser julgada sem que, necessariamente, esses ou outros fatores sejam incorporados às normas dos exames (COE, 2010). Nessa concepção, a *dificuldade* de diferentes exames é vista em termos da probabilidade de sucesso dos candidatos em cada um deles. Dizer que o exame A é "mais difícil" que o exame B não significa fazer referência às exigências de cada um dos exames – como as competências e habilidades necessárias para o sucesso nesses exames –, mas referir-se simplesmente ao fato de que um

aluno *típico* tem maior probabilidade de obter um determinado grau no exame B do que ele teria no exame A. Por aluno *típico* entende-se aquele que pertence à população para a qual o exame foi desenvolvido. Nesse contexto, *típico* pode significar *comparável*, no sentido de que os elementos do grupo possuem um determinado conjunto de características. Desse modo, o exame A é mais difícil que o exame B se os indivíduos com as mesmas características tendem a alcançar notas mais baixas nele (COE, 2010).

A limitação fundamental desse método é que ele considera apenas os fatores determinados pelos exames que estão sendo comparados. Entretanto, o desempenho do estudante nesses exames é afetado por uma ampla gama de fatores, como a quantidade de tempo de estudos, a qualidade do ensino, a motivação dos alunos para os conteúdos verificados nos exames, entre outros, e, a menos que todos esses fatores sejam controlados de forma eficaz, as análises estatísticas podem gerar conclusões equivocadas (COE, 2010; NEWTON, 2007).

Na concepção de *comparabilidade de construto*, deve ser possível relacionar o construto que partilham, pelo menos até certo ponto. As notas provenientes de diferentes exames podem ser interpretadas como indicando o desempenho do participante em relação ao mesmo traço latente ou construto (COE, 2010).

De acordo com a concepção de *comparabilidade de construto*, a *dificuldade* refere-se ao nível de desempenho alcançado em relação ao construto avaliado. Por exemplo, em vez de dizer que o exame A é "mais difícil" do que o exame B, deve-se dizer que um determinado grau no exame A indica um maior nível de habilidade do que esse mesmo grau indicaria no exame B (COE, 2010).

Essa ideia é um conceito central da Teoria de Resposta ao Item (TRI), cuja técnica é utilizada em muitas nações em exames em larga escala (ANDRADE; TAVARES; VALLE, 2000). Dois exemplos importantes de aplicação da TRI em avaliações em larga escala no Brasil, mas somente com itens de múltipla escolha, são o ENEM e o SAEB.

As avaliações sobre a habilidade da expressão escrita e as diversas formas de comparação entre essas avaliações têm recebido muita atenção. Entre os estudos que abordam essas avaliações, estão os que tratam da dificuldade de tarefas utilizadas em testes de proficiência em inglês como segunda língua (HAMP-LYONS; MATHIAS, 1994), da determinação do grau de dificuldade em uma tarefa de escrita e da confiabilidade da pontuação (SUDWEEKS; REEVE; BRADSHAW, 2005) e o estudo sobre a validade e a generabilidade do tema em teste de inglês como segunda língua (LEE; ANDERSON, 2008; PAGANO et al., 2008). Esses são apenas alguns exemplos, embora seja possível encontrar estudos com uma grande variedade de finalidades e enfoques distintos sobre a comparabilidade de avaliações da escrita. No entanto, as provas de redação dos exames vestibulares e do ENEM podem ser consideradas as principais avaliações da escrita em larga escala do Brasil. Mesmo assim, a maioria desses exames é cercada de segredos, e pouco se sabe sobre os processos envolvidos e que resultam em um escore final para o candidato. Segundo Vicentini (2011), Scaramucci (2004) e Vianna (2003), deveria haver mais estudos sobre a eficácia dessas avaliações brasileiras.

Embora a comparação entre avaliações com itens abertos ainda suscite desconfianças e polêmicas, exemplos das vantagens da comparabilidade de testes são comuns em todo o mundo.

Na Europa, nos anos 1990, houve o desenvolvimento de um número significativo de quadros comuns de referência[4] para orientar os currículos e promover uma espécie de *perfil comum* da aprendizagem, tanto no âmbito institucional, como nacional, ou mesmo internacional. Além disso, a natureza multilíngue da população pertencente à União Europeia e o intercâmbio de pessoas motivadas por trabalho ou estudos levaram ao desenvolvimento de um

4- Exemplos de utilização de quadros comuns de referência: University of Cambridge Local Examinations Syndicate (UCLES) e Association of Language Testers in Europe (ALTE).

quadro europeu comum de referência para as línguas europeias (HAMP-LYONS, 2011, 2004).

Nos Estados Unidos, apesar de haver inúmeros exemplos de aplicação e estudos sobre comparabilidade, pesquisas recentes destacam um anseio dos pesquisadores para maior uniformidade nas normas acadêmicas no país (JEFFERY, 2009).

A comparabilidade está presente no sistema de ensino britânico desde o início do século XX, mas, nas últimas décadas, tem-se buscado formas de garantir que os padrões de ensino para a educação básica e também de seleção para as universidades se mantenham constantes de ano para ano. À medida que os exames se tornaram mais competitivos, as exigências de consistência nas pontuações e de comparabilidade provocaram, na Inglaterra, uma série de iniciativas, como um sistema nacional de currículo e avaliação, a utilização de exames em larga escala para avaliar a eficácia dos sistemas de ensinos, a regulação do sistema de ensino, entre outras. A modernização do sistema de exames da Inglaterra está sendo desenvolvida e a comparabilidade consta como um requisito essencial (TATTERSALL, 2007).

Para atender a essa demanda por comparabilidade, muitos trabalhos estão sendo desenvolvidos, visando a uniformizar o significado de comparabilidade e da dificuldade dos exames (ELLIOTT, 2013; BAIRD, 2007; NEWTON, 2007, 2008; POLLITT, AHMED; CRISP, 2007), os métodos estatísticos (ELLIOT, 2013), as normas e padrões comuns (COE, 2010; BAIRD, 2007), as metodologias e as definições sobre comparabilidade (TATTERSALL, 2007; ELLIOTT, 2011), entre outros exemplos.

O Brasil, apesar de timidamente, também tem acompanhado essa tendência mundial por comparações entre resultados das avaliações educacionais em larga escala em diversas áreas. Entre os estudos pioneiros estão Bessa e Mettel (1965), que procuraram estabelecer comparações entre as notas obtidas em exames vestibulares e os escores obtidos previamente pelo grupo de candidatos nos testes do Differential Aptitude Tests (DAT – adaptação do ISOP-FGV), que eram utilizados na época para orientação educacional. Bessa (1980) estudou a associação entre as notas de um grupo de candidatos ao exame vestibular com o desempenho desse grupo no curso de Engenharia. Silveira (1996) analisou as correlações existentes entre avaliações elaboradas com itens objetivos e itens abertos utilizados no exame do vestibular da Fundação Universitária para o Vestibular (FUVEST), que contavam com testes de itens objetivos na primeira fase e testes com itens abertos na segunda fase. Silveira e Pinnent (2001) pesquisaram as correlações entre provas de admissão a duas universidades às quais um mesmo grupo de candidatos foi submetido à mesma época.

Além de pesquisas sobre os exames vestibulares, também existem trabalhos sobre comparações diversas envolvendo outras avaliações em larga escala nacionais. Andrade (2001) elaborou um trabalho no qual aplica a TRI para estabelecer três formas de equalização (ligação) de diferentes populações para obtenção dos parâmetros dos itens e da habilidade dos participantes em uma mesma escala. Fez estudos de simulação e também uma aplicação prática dos resultados na análise do Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo (SARESP) dos anos de 1996 e 1997. Bonamino, Coscarelli e Franco (2002) realizaram uma comparação das habilidades de leitura que foram avaliadas nas provas do SAEB edição de 1999 e do PISA edição de 2000. Klein et al. (2008) fazem comparações entre os resultados dos desempenhos em Língua Portuguesa e em Matemática dos alunos da 4ª e 8ª séries do ensino fundamental e da 3ª série do ensino médio das escolas mantidas pela Fundação Bradesco com resultados provenientes das avaliações Prova Brasil e SAEB.

Outros exemplos de estudos brasileiros nessa linha de pesquisa podem ser encontrados nas principais bases de buscas. Entretanto, percebe-se a inferioridade quanto ao número e quanto à qualidade de pesquisas em

comparação com os trabalhos desenvolvidos no exterior. Na prática, a qualidade das avaliações em larga escala só deve ser alcançada se aliada ao desenvolvimento científico.

Justiça

Uma avaliação de qualidade deve permitir aos participantes condições de respostas que assegurem inferências corretas sobre seu desempenho em relação ao construto medido. As questões sobre justiça estão relacionadas com a equidade da avaliação ou a possibilidade de garantir aos participantes oportunidades iguais. Para isso, os instrumentos devem ser apropriados para os vários grupos que serão testados.

A American Educational Research Association (AERA) e a American Psychological Association (APA) estabelecem que todos os examinandos devem ter oportunidade de demonstrar a sua proficiência em relação ao construto medido pelo teste. A validade do teste depende dessa oportunidade que, por sua vez, depende principalmente dos itens.

Há princípios estabelecidos que se destinam a auxiliar as pessoas responsáveis pelo desenvolvimento de avaliações a entenderem melhor o conceito de equidade na avaliação, evitando-se, desse modo, a inclusão de conteúdos ou imagens que possam provocar injustiças aos participantes ou abordagens de temas que suscitem polêmicas ou que possuam conteúdos considerados sexistas, racistas, ofensivos ou inapropriados (BRASIL, 2010; ETS, 2009).

É necessária, também, a preocupação com grupos de examinandos considerados minoritários, como os raciais, os deficientes visuais e auditivos, e os de pessoas idosas. Esses grupos são significativamente mais afetados por fontes de variância construto-irrelevante do que a população-alvo, o que diminui a justiça, a equidade e a validade do teste. As fontes de variância construto-irrelevante são identificadas por afetar pessoas diferentes de formas diferentes (ETS, 2009).

Deficiências podem prejudicar a capacidade da pessoa de demonstrar seu conhecimento sobre algum construto avaliado. Assim, testes não projetados com a inclusão em mente não podem distinguir adequadamente os que possuem determinada habilidade, mas que são afetados por características do teste, daqueles que simplesmente não possuem tal habilidade. Portanto, deve-se buscar melhorar a validade dos testes e promover a justiça para todos os participantes das avaliações em larga escala (JOHNSTONE et al., 2008).

São muitos os fatores que podem prejudicar a justiça nas avaliações em larga escala. Por isso, são necessárias mais investigações para determinar o modo como os diferentes elementos de um instrumento de avaliação afetam os participantes.

Discussão e conclusão

Nas avaliações educacionais, os objetivos podem estar direcionados para as diferenças individuais, avaliando o desempenho dos estudantes em diversas situações, como também para a avaliação de programas ou de projetos educacionais, subsidiando ou justificando alguma ação na esfera política. Não se pode deixar de destacar o efeito retroativo das avaliações, ou seja, seu impacto no ensino e quanto tal impacto pode servir de guia para a instrução em sala de aula (SCARAMUCCI, 2004, 2011; QUEVEDO-CAMARGO, 2014; ALDERSON; WALL 1993). Desse modo, a validade das medidas e suas interpretações são de suma importância, com consequências para a população envolvida e até mesmo a sociedade. O desenvolvimento de novas metodologias de medição e avaliação que resultem em medidas de maior precisão torna-se cada vez mais importante (HAMP-LYONS, 2002, 2011).

Neste trabalho, procurou-se fornecer uma visão geral das tendências sobre os conceitos de validade, confiabilidade, comparabilidade e justiça nas avaliações em larga escala. No entanto, pesquisadores da área de avaliação

reconhecem que não há metodologia simples que assegure a qualidade da avaliação.

Atualmente, o conceito de validade proposto por Messick (1989) é amplamente aceito. Apesar disso, os aspectos práticos para a determinação da validade e da confiabilidade dos processos avaliativos ainda são problemáticos e, segundo os pesquisadores da área, não possuem solução simples.

McNamara (2000) caracteriza a validade como uma avaliação da própria avaliação e a define como o processo para investigar os procedimentos pelos quais decisões são tomadas a partir das inferências feitas sobre os resultados do teste:

> A validação de um teste envolve o pensar na lógica do teste, especialmente em seu design e em suas intenções, e também envolve olhar para as evidências empíricas – os fatos – que emergem dos dados advindos de um julgamento do teste ou de administrações operacionais. Se não houver procedimentos de validação disponíveis, há potencial para parcialidades e injustiças. Esse potencial é significativo em proporção ao que está em jogo[5] (McNAMARA, 2000, p. 48, tradução nossa).

As inferências sobre os resultados da avaliação frequentemente vão além dos desempenhos observados. Os resultados dos testes não são utilizados simplesmente para relatar como um indivíduo se saiu ao responder alguns itens em determinado momento e sob certas condições. Ao contrário, são usados para apoiar afirmações diversas, como afirmar que um indivíduo possui certo nível de habilidade em algum construto ou alguma probabilidade de sucesso em um programa educacional. Essas afirmações geralmente não são evidentes nas avaliações. É necessário examinar a plausibilidade das afirmações com base nos resultados dos testes para validar as interpretações e utilizações desses resultados (KANE, 2012).

Atualmente, há uma forte tendência para que as avaliações estejam mais direcionadas para a avaliação da aprendizagem, no lugar dos testes tradicionais de conhecimentos, o que tem intensificado o interesse pelas avaliações de desempenho. Acredita-se que os testes com itens de respostas abertas podem reproduzir atividades relacionadas ao mundo real do estudante, já que a aprendizagem é um produto do contexto em que ocorre (JONSSON; SVINGBY, 2007; MESSICK, 1996). Assim, esse tipo de avaliação pode refletir melhor a complexidade da realidade e fornecer dados mais válidos sobre a competência da pessoa que está sendo avaliada (DARLING-HAMMOND; SNYDER, 2000).

Nas avaliações de desempenho, são muitos os fatores que podem afetar a medida do desempenho das pessoas ao executar a tarefa determinada no teste. Em primeiro lugar está a habilidade do examinando, mas a pontuação que ele receberá no exame não depende apenas da sua capacidade ou do conhecimento sobre o construto sendo medido. Depende também da severidade do avaliador, da dificuldade das tarefas, do formato da questão, do tema abordado, dos critérios e da escala de pontuação e de outras variáveis que podem interferir em cada evento de avaliação em particular.

Esses e outros fatores são frequentemente constatados em estudos relacionados com avaliações com itens abertos, principalmente nas avaliações da linguagem de primeira e segunda língua. Alguns exemplos podem ser obtidos nos trabalhos de Huang (2012), Gyagenda e Engelhard (2009), Jonsson e Svigby (2007), Sudweeks, Reeve e Bradshaw (2005). Entretanto, são poucas as pesquisas relatando resultados de avaliações em larga escala brasileiras, especialmente sobre as avaliações com itens abertos. Em se tratando das redações dos principais exames vestibulares e do ENEM, essa escassez é ainda maior. A maioria dos trabalhos ocorre na área

5- No original em inglês: Test validation similarly involves thinking about the logic of the test, particularly its design and its intentions, and also involves looking at empirical evidence – the hard facts – emerging from data from test trials or operational administrations. If no validation procedures are available there is potential for unfairness and injustice. This potential is significant in proportion to what is at stake (McNAMARA, 2000, p. 48).

da linguística aplicada. Entretanto, são raras as pesquisas sobre a validade e a confiabilidade das avaliações, principalmente as proporcionadas pelas instituições promotoras das avaliações. Sobre a redação do ENEM, as divulgações mais frequentes são as notícias da ocorrência de problemas relatados pela imprensa.

As exigências por avaliações desenvolvidas segundo padrões atuais de qualidade são primordiais. É necessário que as empresas provedoras de tais exames tenham conhecimento, atendam a esses padrões e comprovem a qualidade de seus exames com monitoramentos constantes e pesquisas publicadas.

Referências

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION (AERA); AMERICAN PSYCHOLOGICAL ASSOCIATION (APA); NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (NCME). **Standards for educational and psychological testing.** Washington, DC: American Educational Research Association, 1999.

ANDRADE, Dalton Francisco. Comparando desempenhos de grupos de alunos por intermédio da Teoria da Resposta ao Item. **Estudos em Avaliação Educacional**, São Paulo, n. 23, p. 31-69, 2001.

ANDRADE, Dalton Francisco; TAVARES, Heliton Riberiro; VALLE, Raquel da Cunha. **Teoria da resposta ao item:** conceitos e aplicações. São Paulo: Associação Brasileira de Estatística, 2000.

ALDERSON, J. Charles; WALL, Dianne. Does washback exist? **Applied Linguistics,** Oxford, v. 14, n. 2, p. 115-129, 1993.

BAIRD, Jo-Anne. Alternative conceptions of comparability. In: NEWTON, Paul et al. (Ed.).**Techniques for monitoring the comparability of examination standards.** London: QCA, 2007. p. 124-165.

BEHIZADEH, Nadia; ENGELHARD, Georg Jr. Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. **Assessing Writing,** v. 16, n. 3, p. 189-211, 2011.

BESSA, Nicia Maria. Aspectos metodológicos do processo de seleção para o ingresso nas universidades. **Educação e Seleção,** São Paulo, n. 2, p. 39-56, dez. 1980.

BESSA, Nicia Maria. Validade: o conceito, a pesquisa, os problemas de provas geradas pelo computador. **Estudos em Avaliação Educacional,** São Paulo, v. 18, n. 37, p. 115-156, 2007.

BESSA, Nicia Maria; METTEL, Thereza Lemos. Validade de três testes do DAT (Forma B). **Arquivos Brasileiros de Psicotécnica,** Rio de Janeiro, v. 14, n. 3, p. 5-15, 1965.

BONAMINO, Alicia; COSCARELLI, Carla; FRANCO, Creso. Avaliação e letramento: concepções de aluno letrado subjacentes ao SAEB e ao PISA. **Educação e Sociedade,** Campinas, v. 23, n. 81, p. 91-113, 2002.

BORSBOOM, Denny; MELLENBERG, Gideon J.; VAN HEERDEN, Jaap. The concept of validity. **Psychological Review,** Washington, DC, v. 111, n. 4, p. 1061-1071, 2004.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). **A redação no Enem 2013:** guia do participante. Brasília, DF: MEC, 2013. Disponível em: <http://download.inep.gov.br/educacao_basica/enem/guia_participante/ 2013/guia_participante_redacao_ enem_2013.pdf>. Acesso em: 6 set. 2013.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). **Guia de elaboração e revisão de itens.** v. 1. Brasília, DF: MEC, 2010. Disponível em: <http://download.inep.gov.br/outras_acoes/bni/guia/guia_elaboracao_ revisao_itens_2012.pdf>. Acesso em: 22 out. 2013.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). **Programme for International Student Assessment** (Pisa) - Programa Internacional de Avaliação de Estudantes. Brasília, DF: MEC, 2011. Disponível em: <http://portal.inep.gov.br/pisa-programa-internacional-de-avaliacao-de-alunos>. Acesso em: 20 jun. 2014.

CHAPELLE, Carol A. Validity in language assessment. **Annual Review of Applied Linguistics,** Cambridge, v. 19, p. 254-272, 1999.

COE, Robert. Understanding comparability of examination standards. **Research Papers in Education,** v. 25, n. 3, p. 271–284, 2010.

DARLING-HAMMOND, Linda; SNYDER, Jon. Authentic assessment of teaching in context. **Teaching and Teacher Education,** n. 16, n. 5-6, p. 523-545, 2000.

ECKES, Thomas. **Introduction to many-facet rasch measurement:** analyzing and evaluating rater-mediated assessment. Frankfurt: Peter Lang, 2011.

ELLIOTT, Gill. A guide to comparability terminology and methods. **Research Matters,** Cambridge, special issue 2, p. 9-19, 2013.

ENGELHARD, Georg Jr. **Invariant measurement:** using Rasch models in the social, behavioral, and health sciences. New York: Routledge Academic, 2013.

ETS **International principles for fairness review of assessments:** a manual for developing locally appropriate fairness review guidelines in various Countries. Princeton: Educational Testing Service, 2009.

FONTANIVE, Nilma, et al. A alfabetização de crianças de 1° e 2° ano do ensino fundamental de 9 anos: uma contribuição para a definição de uma matriz de competências e habilidades de leitura, escrita e matemática. **Ensaio,** Rio de Janeiro, v. 18, n. 68, p. 527-548, 2010.

GUILFORD, Joy Paul. **Psychometric methods.** New York: McGraw-Hil, 1954.

GYAGENDA, Ismail S.; ENGELHARD, Georg Jr. Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. **Journal of Applied Measurement,** v. 10, n. 3, p. 225-246, 2009.

HAERTEL, Edward Henry; LINN, Robert L. Comparability. In: PHILLIPS, Gary (Ed.). **Technical issues in large-scale performance assessment.** Washington, DC: National Center for Education Statistics, 1996. p. 1-18.

HAMP-LYONS, Liz. Writing assessment: expanding outwards and coming together. **Assessing Writing,** v. 13, n. 1, p. 1-3, 2004.

HAMP-LYONS, Liz. Writing assessment: shifting issues, new tools, enduring questions. **Assessing Writing,** v. 16, n. 1, p. 3-5, 2011.

HAMP-LYONS, Liz; MATHIAS, Sheila Prochnow. Examining expert judgments of task dificulty on essay tests. **Journal of Second Language Writing,** v. 3, n. 1, p. 49-68, 1994.

HUANG, Jinyan. Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. **Assessing Writing,** v. 17, n. 3, p. 123-139, 2012.

HUOT, Brian. The literature of direct writing assessment: major concern and prevailing trends. **Review of Educational Research.** Thousand Oaks, v. 60, p. 237-264, 1990.

HUOT, Brian. Toward a new theory of writing assessment. **National Council of Teachers of English,** Urbana, v. 47, n. 4, p. 549-566, 1996.

JEFFERY, Jill V. Constructs of writing proficiency in US state and national writing assessments: exploring variability. **Assessing Writing**, v. 14, n. 1, p. 3-24, 2009.

JONSSON, Anders; SVINGBY, Gunilla. The use of scoring rubrics: reliability, validity and educational consequences. **Educational Research Review,** v. 2, n. 2, p. 130-144, 2007.

JOHNSTONE, Christopher J.; et. al. Universal design and multimethod approaches to item review. **Educational Measurement,** v. 27, n. 1, p. 25-36, 2008.

KANE, Michael. Validating score interpretations and uses. **Language Testing,** v. 29, n. 1, p. 3-17, 2012.

KELLEY, Truman Lee. **Interpretation of educational measurements.** New York: Macmillan, 1927.

KLEIN, Ruben; et al. O desempenho dos alunos da Fundação Bradesco: uma comparação com os resultados do SAEB. **Estudos em Avaliação Educacional,** São Paulo, v. 19, n. 41, p. 499-515, 2008.

KLEIN, Ruben; FONTANIVE, Nilma. Uma nova maneira de avaliar as competências escritoras na redação do ENEM. **Ensaio,** Rio de Janeiro, v. 17, n. 65, p. 585-598, 2009.

LEE, Hee-Kyung; ANDERSON, Carolyn. Validity and topic generality of a writing performance test. **Language Testing,** v. 24, n. 3, p. 307-330, 2008.

LI, Heng. The resolution of some paradoxes related to reliability and validity. **Journal of Educational and Behavioral Statistics,** Thousand Oaks, v. 28, n. 2, p. 89-95, 2003.

LORD, Frederic M.; NOVICK, Melvin R.; BIRNBAUM, Allan. **Statistical theories of mental test scores.** Massachusetts: Addison Weslley. 1968.

McNAMARA, Tim. **Language testing.** Oxford: Oxford University Press, 2000.

MESSICK, Samuel. Validity of performance assessments. In: PHILLIPS, Gary W. (Ed.). **Technical issues in large-scale performance assessment.** Washington, DC: National Center for Education Statistics, 1996. p. 1-18.

MESSICK, Samuel. Validity. In: LINN, Robert L. (Ed.). **Educational measurement.** 3. ed. New York: Macmillan, 1989. p. 13-103.

MOSKAL, Barbara M.; LEYDENS, Jon A. Scoring rubric development: validity and reliability. **Practical Assessment, Research & Evaluation,** v. 7, n. 10, p. 71-81, 2000. Disponível em: <http://pareonline.net/getvn.asp?v=7&n=10>. Acesso em: 3 fev. 2013.

MYFORD, Carol M.; WOLFE, Edward W. Detecting and measuring rater effects using many-facet rasch measurement: part II. **Journal of Applied Measurement,** v. 5, n. 2, p. 189-227, 2004.

NEWTON, Paul. Comparability monitoring: progress report. In: NEWTON, Paul et al. (Ed.).**Techniques for monitoring the comparability of examination standards.** London: QCA, 2007. p. 452-476.

PASQUALI, Luiz. Validade dos testes psicológicos: será possível reencontrar o caminho? **Psicologia:** Teoria e Pesquisa. Brasília, DF, v. 23, n. especial, p. 99-107, 2007.

PAGANO, Neil et al. An inter-institutional model for college writing assessment. **College Composition and Communication,** Illinois, v. 60, n. 2, p. 285-320, 2008.

POLLITT, Alastair; AHMED, Ayesha.; CRISP, Victoria. The demands of examination syllabuses and question papers. In: NEWTON, Paul et al. (Ed.).**Techniques for monitoring the comparability of examination standards.** London: QCA, 2007. p. 166-206. London: QCA, 2007.

POMPLUN, Mark; et al. An analysis of English composition test essay prompts for differential difficulty. **College Board Report,** New York: College Entrance Examination Board, 1992.

QUEVEDO-CAMARGO, Gladys. Efeito retroativo da avaliação na aprendizagem de línguas estrangeiras: que fenômeno é esse? In: MULIK, Katia Bruginsk; RETORTA, Miriam Sester (Org.). **Avaliação no ensino-aprendizado de línguas estrangeiras:** diálogos, pesquisas e reflexões. v. 1. Campinas: Pontes, 2014. p. 1-16.

SCARAMUCCI, Matilde Virginia Ricard. Efeito retroativo da avaliação no ensino/aprendizagem de línguas: o estado da arte. **Trabalhos em Linguística Aplicada,** Campinas, v. 2, n. 43, p. 203-226, 2004.

SCARAMUCCI, Matilde Virginia Ricard. Validade e consequências sociais das avaliações em contexto de ensino de línguas. **Lingvarvm Arena,** Porto, v. 2, p. 103-120, 2011.

SISTEMA NACIONAL DE AVALIAÇÃO DA EDUCAÇÃO BÁSICA (SAEB). Disponível em: <http://www.inep.gov.br/download/saeb/1999/resultados/escalas_saeb.DOC>. [e] <http://www.inep.gov.br/saeb/>. Acesso em: 14 abr. 2013.

SILVEIRA, Fernando Lang. Correlação entre avaliações por testes de múltipla escolha e por provas analítico-expositivas. **Revista Brasileira de Ensino de Física,** São Paulo, v. 18, n. 4, p. 362-371, 1996.

SILVEIRA, Fernando Lang; PINNENT, Carlos Eduardo. A questão da redação no concurso vestibular à universidade: validade e poder decisório. **Estudos em Avaliação Educacional,** São Paulo, n. 24, p. 147-164, 2001.

SLOMP, David H.; FUITE, Jim. Following Phaedrus: Alternate choices in surmounting the reliability/validity dilemma. **Assessing Writing,** n. 9, n. 3, p. 190-207, 2005.

SUDWEEKS, Richard R.; REEVE, Suzanne; BRADSHAW, William S. A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. **Assessing Writing,** v. 9, n. 3, p. 239-261, 2005.

TATTERSALL, Kathleen. A brief history of policies, practices and issues relating to comparability. In: NEWTON, Paul et al. (Ed.).**Techniques for monitoring the comparability of examination standards.** London: QCA, 2007. p. 43-96.

VIANNA, Heraldo Merelim. Avaliações nacionais em larga escala: análises e propostas. **Estudos em Avaliação Educacional,** São Paulo, n. 27, p. 41-76, 2003.

VICENTINI, Monica Panigassi. **Exame nacional do ensino médio:** a relevância de pesquisas empíricas sobre validade e efeitos retroativos. 2011. Dissertação (Mestrado) - Universidade Estadual de Campinas, Campinas, 2011.

YANCEY, Kathleen Blake. Looking back as we look forward: historicizing writing assessment. College Composition and Communication, **Urbana,** v. 50, n. 3, p. 483-503, 1999.

Recebido em: 05.06.2014

Aprovado em: 10.03.2015

Sônia Ferreira Lopes Toffoli é mestre em matemática aplicada pela Universidade Estadual de Campinas (UNICAMP), doutora em engenharia de produção pela Universidade Federal de Santa Catarina e professora adjunta da Universidade Estadual de Londrina.

Dalton Francisco de Andrade é professor doutor titular aposentado do Departamento de Informática e Estatística da Universidade Federal de Santa Catarina. Atualmente, é professor voluntário junto aos programas de pós-graduação do Departamento de Engenharia de Produção (PPGEP), e do Programa de Pós-Graduação em Métodos e Gestão em Avaliação (PPGMGA), do Departamento de Informática e Estatística, ambos da Universidade Federal de Santa Catarina, e pesquisador associado da Fundação Vunesp.

Antonio Cezar Bornia é professor titular do Departamento de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina. Possui graduação em engenharia mecânica pela Universidade Federal do Paraná (1985), mestrado em engenharia de produção pela Universidade Federal de Santa Catarina (1988) e doutorado em engenharia de produção pela Universidade Federal de Santa Catarina (1995).

Gladys Quevedo-Camargo é doutora em estudos da linguagem pela Universidade Estadual de Londrina. Professora do Departamento de Línguas Estrangeiras e Tradução, Universidade de Brasília, Brasília, DF.